Multiple imputation of item scores when test data are factorially complex

van Ginkel, J.R.; van der Ark, L.A.; Sijtsma, K.

Published in:
British Journal of Mathematical and Statistical Psychology

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record

British Journal of Mathematical and Statistical Psychology (2007), 60, 315–337

The British Psychological Society
www.bpsjournals.co.uk

Multiple imputation for item scores when test data are factorially complex

Joost R. van Ginkel*, L. Andries van der Ark and Klaas Sijtsma
Tilburg University, The Netherlands

Multiple imputation under a two-way model with error is a simple and effective method that has been used to handle missing item scores in unidimensional test and questionnaire data. Extensions of this method to multidimensional data are proposed. A simulation study is used to investigate whether these extensions produce biased estimates of important statistics in multidimensional data, and to compare them with lower benchmark listwise deletion, two-way with error and multivariate normal imputation. The new methods produce smaller bias in several psychometrically interesting statistics than the existing methods of two-way with error and multivariate normal imputation. One of these new methods clearly is preferable for handling missing item scores in multidimensional test data.

1. Introduction

This study deals with the imputation of scores in incomplete, multidimensional rating-scale data stemming from questionnaires used in psychological, sociological and other research. Examples of such multidimensional data come from questionnaires intended to measure different ways of being religious (Hills, Francis, & Robbins, 2005), different aspects of schizotypal personality disorder (Mata, Mataix-Cols, & Peralta, 2005), different coping styles (Brough, O'Driscoll, & Kalliath, 2005), and different kinds of phobias (Brown, White, & Barlow, 2005). Each subset of items in such a questionnaire measures one dimension of a broader construct and different subsets measure different dimensions. The data are often collected by means of group, mail, telephone and Internet testing, each of which gives ample rise to the occurrence of missing data.

The focus of this study is item non-response – the respondent leaves at least one answer open but also provides at least one answer so that his/her data record is incomplete but not completely missing. Item non-response may have many causes, such as embarrassment (e.g. invasion of privacy), secrecy (e.g. income, career history), boredom (e.g. too many questions), misunderstanding (e.g. unfortunate phrasing of

*Correspondence should be addressed to Joost R. van Ginkel, VU University Medical Centre, Department of Clinical Epidemiology & Biostatistics, PK 62 183, PO Box 7057, 1007 MB, Amsterdam, The Netherlands (e-mail: j.vanginkel@vumc.nl).

DOI:10.1348/000711006X117574

questions), stubbornness (e.g. reluctance to cooperate) or sloppiness. The multidimensional or multifactor structure in the available data is used to obtain good item-score estimates by means of simple multiple-imputation methods. Hopefully, statistical results based on this *completed* data matrix show little bias and also little *discrepancy* relative to the results based on the original, complete data.

Item scores may be missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) (Little & Rubin, 2002, p. 12; Rubin, 1976). Let N be the number of participants who filled out a questionnaire consisting of J items, and let $\mathbf{X}$ be the resulting $N \times J$ data matrix consisting of scores X_{ij} ($i = 1, \ldots, N; j = 1, \ldots, J$). Furthermore, let $\mathbf{R}$ be a response-indicator matrix with entry $R_{ij} = 1$ if score X_{ij} in $\mathbf{X}$ is observed, and $R_{ij} = 0$ if score X_{ij} in $\mathbf{X}$ is missing. Finally, let $\boldsymbol{\xi}$ be a parameter vector that explains the missingness. MCAR means that the item-score missingness is related neither to the observed part of data matrix $\mathbf{X}$ (denoted $\mathbf{X}_{\text{obs}}$) nor to the unobserved part ($\mathbf{X}_{\text{mis}}$), and is formalized as

$$P(\mathbf{R}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}, \boldsymbol{\xi}) = P(\mathbf{R}|\boldsymbol{\xi}). \tag{1}$$

MAR means that missingness depends on completely observed covariates, so that:

$$P(\mathbf{R}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}, \boldsymbol{\xi}) = P(\mathbf{R}|\mathbf{X}_{\text{obs}}, \boldsymbol{\xi}). \tag{2}$$

NMAR means that the missing item score X_{ij} depends either on variables that were not collected, or on the unobserved value of X_{ij} itself, or both. When the missingness parameters ($\boldsymbol{\xi}$) and the parameters that govern the data are distinct, MAR and MCAR represent ignorable missingness, and NMAR non-ignorable missingness. Otherwise the missingness is always non-ignorable. We assume that the parameters are distinct.

Multiple imputation (MI) is a procedure recommended for handling missing data (Rubin, 1987, p. 9). MI estimates the missing data w times using a stochastic population model, resulting in w different plausible, complete data sets. The results from statistical analyses on these w data sets are combined into one conclusion. Accordingly, uncertainty about missing values is taken into account. Multivariate normal imputation is available in the programs NORM (Schafer, 1998), S-plus 6 for Windows (2001) and SAS 8.1 (Yuan, 2000). S-plus also performs MI under the saturated logistic model and the general location model. These methods produce statistical results with little bias (Ezzati-Rice *et al.*, 1995; Graham & Schafer, 1999; Schafer, 1997; Schafer *et al.*, 1996). Each method assumes ignorable missingness. For non-ignorable missingness methods, see Heckman (1976) for continuous data, and Fay (1986) and O'Muircheartaigh and Moustaki (1999) for categorical data.

Many practical researchers have only been trained in basic data analysis and do not have a statistician available who can help them use relatively complicated methods. They often resort to methods such as listwise deletion that produce biased and less efficient results (Schafer & Graham, 2002). Alternatively, imputation methods may be used, such as two-way imputation (TW; Bernaards & Sijtsma, 2000), corrected item-mean substitution (Huisman, 1998, p. 96), relative mean imputation (Raaijmakers, 1999) and response-function imputation (Sijtsma & Van der Ark, 2003); see also Smits, Mellenbergh, and Vorst (2002).

Van der Ark and Sijtsma (2005) and Van Ginkel, Van der Ark, and Sijtsma (2007) showed that an MI version of method TW produced little discrepancy and little loss of efficiency in Cronbach's (1951) alpha, Loevinger's (1948) H, the item-cluster solution from Mokken (1971) scale analysis, and fit statistics for the Rasch (1960) model, for

unidimensional data and correlated (at least 0.24) two-dimensional data (Bernaards & Sijtsma, 2000; Van Ginkel *et al.*, 2007). This study discusses new versions of method TW that deal explicitly with general forms of multidimensional data by making use of the correlation structure of the items in the questionnaire.

2. Method

2.1. Outline of methodology

In outline, the methodology of this study is as follows:

(1) Complete data sets of multidimensional rating-scale scores were simulated using a multidimensional item response model, producing original data sets. Statistics of interest were estimated from these data. For each combination of questionnaire and population, 100 original data sets were sampled. This enabled estimation of the sampling variation of statistics of interest.
(2) Item scores were deleted from original data sets, thus creating data sets with missing item scores. These were called incomplete data sets.
(3) For an incomplete data set, a multiple-imputation method was used to estimate the missing item scores five times yielding five completed data sets.
(4) The statistical calculations on each of the five completed data sets were combined using Rubin's (1987) rules.
(5) Steps 2, 3 and 4 were repeated for each of the 100 independently sampled original data sets. The mean and the sampling variation were determined of the bias in, for example, Cronbach's alpha. When the complete data produce biased estimates of Cronbach's alpha, the discrepancy in Cronbach's alpha between original data and completed data may be a better indicator of the performance of the multiple-imputation methods. Thus, the mean and sampling variation of this discrepancy measure were also determined. Imputation methods should produce little bias and discrepancy.

2.2. Missing-data methods

2.2.1. Listwise deletion (LD)

Method LD – removal of all incomplete cases prior to analysis – was used as lower benchmark.

2.2.2. Two-way imputation with normally distributed errors (TW-E) – unidimensional-data case

Following ANOVA, method TW-E (Bernaards & Sijtsma, 2000; also, see Little & Su, 1989) imputes scores using a person and an item effect. Let $obs(i)$ be the set of observed item scores of person i, and $\#obs(i)$ the size of this set. The mean of the $\#obs(i)$ observed scores for person i is denoted PM_i. Likewise, item mean IM_j for the observed scores on item j and the overall mean OM for all observed item scores in $\mathbf{X}$ are defined. First, for cell (i, j), in which $R_{ij} = 0$, define an expected item score, denoted TW_{ij}, as

$$TW_{ij} = PM_i + IM_j - OM. \tag{3}$$

Second, let ε_{ij} be a random error from $N(0, \sigma^2_\varepsilon)$. To estimate σ^2_ε, expected item scores are computed for all observed scores using equation (3). This results in estimate

$$S^2_\varepsilon = \sum_{i,\, j \in obs} \sum (X_{ij} - TW_{ij})^2 / (\#obs - 1).$$

Third, ε_{ij} is drawn from $N(0, S_{\varepsilon}^2)$, and added to TW_{ij}, so that

$$TW_{ij}(E) = TW_{ij} + \varepsilon_{ij}.$$

Let item scores be adjacent, ordered integers, denoted $x_{\min}, \ldots, x_{\max}$, and round $TW_{ij}(E)$ to the nearest feasible integer. The result is imputed in cell (i, j).

2.2.3. Two-way imputation – multidimensional-data case

Main types of two-way imputation

Throughout this section, we use factor loadings resulting from principal components analysis followed by varimax rotation (PCA/VR) for weighting item scores. PCA/VR is often used for determining the dimensionality of questionnaire data. We assume that the number of principal components was equal to the number of dimensions in the simulated data. Table 1 shows six new extensions of method TW to multidimensional data. They represent two main types with three variations each.

Table 1. Missing-data methods

Missing-data method	Factor loadings obtained from:	PM_i and OM are computed using:	Are item scores weighted with loadings?
TW-SS_{tw}	Completed data using TW-E	Only items in scale k	No
TW-SS_{bs}	Bootstrap sample	Only items in scale k	No
TW-SS_{od}	Original data	Only items in scale k	No
TW-FL_{tw}	Completed data using TW-E	All items	Yes
TW-FL_{bs}	Bootstrap sample	All items	Yes
TW-FL_{od}	Original data	All items	Yes
TW-E	–	All Items	No
LD	–	–	–
MNI	–	–	–

Main type I: Two-way imputation for separate scales (TW-SS). Assume the availability of a PCA/VR solution for the incomplete data matrix **X** (how this solution is obtained, will be discussed shortly). Next, apply method TW-E separately to each item subset consisting of the items that load highest on the same rotated factor. This main type is denoted TW-SS ('SS' for separate scales).

Main type II: Two-way with factor loadings (TW-FL). Assume that $R_{ij} = 0$ and that item j has its highest loading on the kth rotated factor; denote this loading a_{jk}. Method TW-FL ('FL' for factor loadings) uses a different estimate of the person mean than methods TW-E and TW-SS, and weights available item scores with the items' loadings on factor k. As a point of departure, consider

$$PM_{ik}^{\#} = \frac{\sum_{j \in \mathrm{obs}(i)} a_{jk} \times X_{ij}}{\sum_{j \in \mathrm{obs}(i)} a_{jk}}. \tag{4}$$

Note that negative loadings may have the effect of reducing the denominator so that the behaviour of $PM_{ik}^{\#}$ is unpredictable. This effect can be corrected by means of the

midrange score of item j, defined as

$$x_{\text{mid}} = \frac{1}{2}(x_{\max} - x_{\min}),$$

and, using x_{mid}, by defining an alternative person mean as

$$PM^*_{ik} = \frac{\sum_{j \in obs(i)} a_{jk} \times (X_{ij} - x_{\text{mid}})}{\sum_{j \in obs(i)} |a_{jk}|} + x_{\text{mid}}. \tag{5}$$

This definition does not suffer from the undesirable effect that negative loadings may have; see the Appendix for details. The item means and the overall means can be defined similarly, but technical details are ignored here.

Using these corrected means, the expected value $TW^*_{ij,\,k}$ is computed as

$$TW^*_{ij,\,k} = IM^*_{jk} + PM^*_{ik} - OM^*_{k}. \tag{6}$$

Let $obs(k)$ be the set of all observed scores on the items that load highest on factor k, and let $\#obs(k)$ be the size of this set. The error variance of these data is estimated as

$$S^{2*}_{\varepsilon,\,k} = \sum_{i,\,j \in obs(k)} \sum (X_{ij} - TW^*_{ij,\,k})^2 / [\#obs(k) - 1].$$

For cell (i, j) we obtain

$$TW_{ij,\,k}(FL) = TW^*_{ij,\,k} + \varepsilon_{ij,\,k},$$

with $\varepsilon_{ij,\,k} \sim N(0, S^{2*}_{\varepsilon,\,k})$. Finally, $TW_{ij,k}(FL)$ is rounded to the nearest feasible integer and the result is imputed in cell (i, j).

Specific types of two-way imputation

We distinguish three versions of method two-way for separate scales (Main type I):

(1) *Method TW-SS using two-way (TW-SS$_{tw}$).* Method TW-SS$_{\text{tw}}$ has the following steps: (1) Item scores are imputed in the incomplete data using method TW-E, ignoring the dimensionality of the data; (2) PCA/VR is applied to the completed data set, and item subsets are identified by the items' highest loadings; and (3) item scores are imputed anew in the incomplete data using method TW-E, but now for each item subset separately. This process is repeated five times yielding five completed data sets.

(2) *Method TW-SS using bootstrap sampling (TW-SS$_{bs}$).* Method TW-SS$_{\text{tw}}$ does not propagate error in the factor loadings; thus it is improper (Rubin, 1987). A reviewer suggested remedying this by means of bootstrap sampling: (1) A bootstrap sample is drawn from the incomplete data set; (2) method TW-E is applied to this bootstrap sample; (3) PCA/VR is applied to the completed data; and (4) method TW-SS is applied to the incomplete data set, using factor loadings obtained from the completed bootstrap data set. This process is repeated five times. Method TW-SS$_{\text{bs}}$ is a refinement of method TW-SS$_{\text{tw}}$, and was studied in a specialized design.

(3) *Method TW-SS using original data (TW-SS$_{od}$).* PCA/VR on the original data yielded factor loadings that could be used for identifying item subsets and, using this information method TW-SS could be used for imputing scores in the incomplete

data. This is method TW-SS_{od}. Five data sets are created. This is not a practically useful method, but it provided information on the amount of discrepancy produced by method TW-SS_{tw} due to using factor loadings obtained from a completed data set in which scores were imputed using method TW-E, and method TW-SS_{bs} due to using factor loadings obtained from a completed bootstrap data set in which scores were imputed using method TW-E.

Also, three versions of method two-way using factor loadings (Main type II) are distinguished:

(1) *Method TW-FL using two-way (TW-FL_{tw}).* Method TW-FL_{tw} has the following steps: (1) Item scores are imputed in the incomplete data using method TW-E, ignoring the dimensionality of the data; (2) the PCA/VR solution for this completed data set is used to identify item subsets; and (3) method TW-FL is used to impute item scores in the incomplete data using the factor loadings found in the second step. This is repeated to obtain five completed data sets.
(2) *Method TW-SS using bootstrap sampling (TW-SS_{bs}).* Method TW-FL_{bs} has the following steps. (1) A bootstrap sample is drawn from the incomplete data; (2) this bootstrap sample is completed using method TW-E; (3) PCA/VR is applied to this completed bootstrap data set; and (4) the resulting factor solution is used for imputation with method TW-FL_{bs}. This is repeated five times. Method TW-FL_{bs} was studied in a specialized design.
(3) *Method TW-FL using original data (TW-FL_{od}).* This method has the following steps: (1) Method TW-FL_{od} uses the PCA/VR solution for the original data; and (2) method TW-E is applied to each of the item subsets resulting from PCA/VR. Five data sets are created. This method cannot be used in practice but is used to assess discrepancy compared with methods TW-FL_{tw} and TW-FL_{bs}.

2.2.4. Multivariate normal imputation (MNI)

Method MNI (Schafer, 1998) uses data augmentation (Tanner & Wong, 1987), which obtains the distribution of the missing item scores, given the observed data. Scores are imputed by random draws from the multivariate normal distribution. Starting values are obtained using an EM algorithm (Dempster, Laird, & Rubin, 1977). In this study, imputed scores were rounded to the nearest integer within the range $x_{\min}, \ldots, x_{\max}$.

MI under the saturated logistic model could have been a more natural choice, but its application was found to be problematic for large data sets (cf. Van der Ark & Sijtsma, 2005). Also, Ezzati-Rice *et al.* (1995) and Schafer *et al.* (1996) showed that MNI is robust to departures from the multivariate normal model. MNI is available in the programs NORM, S-plus 6 and SAS 8.01.

2.3. Setup of simulation study

2.3.1. Fixed design characteristics

Item scores. Data sets were simulated using the multidimensional polytomous latent trait (MPLT) model (Kelderman & Rijkes, 1994). Four latent variables were assumed, denoted θ_q ($q = 1, \ldots, Q$; here $Q = 4$). Parameter θ_{iq} is the value of person i on latent variable q. Also, ψ_{jqx} is the separation parameter of item j for latent variable q and answer category x; and B_{jqx}($B_{jqx} \geq 0$) is the discrimination parameter of item j with respect to

latent variable q and score x. The MPLT model is defined as

$$P(X_{ij} = x|\theta_{i1}, \ldots, \theta_{iQ}) = \frac{\exp\left[\sum_{q=1}^{Q}(\theta_{iq} - \psi_{jqx})B_{jqx}\right]}{\sum_{y=0}^{x}\left\{\exp\left[\sum_{q=1}^{Q}(\theta_{iq} - \psi_{jqy})B_{jqy}\right]\right\}}. \quad (7)$$

Parameters B_{jq0} and ψ_{jq0} must be set to 0 to ensure uniqueness of the parameters. Data sets of 40 polytomously scored items with five answer categories (i.e. $x_{\min} = 0$, $x_{\max} = 4$) were simulated. Items 1–10 were driven by θ_1, items 11–20 by θ_2, items 21–30 by θ_3 and items 31–40 by θ_4.

Item parameters. Table 2 shows the item parameters (based on Van Ginkel *et al.*, 2007). Items with an even index in the range 1–20 had $B_{jqx} = 2$ (i.e. high discrimination) and items with an odd index had $B_{jqx} = 0.5$ (i.e. low discrimination). For items 21–40 this was reversed. The separation parameters ψ_{jqx} ranged from -2.75 to 2.75. For each latent variable, item difficulty increased with increasing item index.

Table 2. Location parameters ψ_{jqx} and discrimination parameters B_{jqx} of simulated data

Items	ψ_{jq1}	ψ_{jq2}	ψ_{jq3}	ψ_{jq4}	B_{jqx}
1, 11, 22, 32	1.25	1.75	2.25	2.75	0.5
2, 12, 21, 31	1.25	1.75	2.25	2.75	2
3, 13, 24, 34	0.25	0.75	1.25	1.75	0.5
4, 14, 23, 33	0.25	0.75	1.25	1.75	2
5, 15, 26, 36	−0.75	−0.25	0.25	0.75	0.5
6, 16, 25, 35	−0.75	−0.25	0.25	0.75	2
7, 17, 28, 38	−1.75	−1.25	−0.75	−0.25	0.5
8, 18, 27, 37	−1.75	−1.25	−0.75	−0.25	2
9, 19, 30, 40	−2.75	−2.25	−1.75	−1.25	0.5
10, 20, 29, 39	−2.75	−2.25	−1.75	−1.25	2

Covariate classes. Dichotomous covariate Y was used for simulating missingness mechanisms. For $Y = 1$, four latent variable values were randomly drawn from a multivariate normal distribution with $\boldsymbol{\mu}_1 = [-.25, -.25, -.25, -.25]$. Likewise, for $Y = 2$ we used $\boldsymbol{\mu}_2 = [.25, .25, .25, .25]$. Both covariance matrices of the latent variables were equal to express that the items measured the same constructs in, for example, gender groups. Covariance matrices equaled the correlation matrices, with ones on the main diagonal and elements ρ on the off-diagonal places.

2.3.2. Independent variables

Sample size. The sample sizes $N = 300$ ('fair') and $N = 1000$ ('excellent') were based on rules of thumb for PCA (Comrey & Lee, 1992).

Correlation between latent variables. The correlation (ρ) between the latent variables was varied to be 0, .24 and .50 (see Bernaards & Sijtsma, 2000).

Percentage of missing item scores. One, 5 and 15% missing item scores were simulated.

Effect of covariate on missingness. For missingness unrelated to *Y*, the probability of scores being missing was equal for both classes. For missingness related to *Y*, the probability of scores being missing was twice as high for $Y = 2$ as for $Y = 1$. Given these relative probabilities, a random sample of item scores was removed from the original data matrix.

Joint effect of item score and item location parameters on missingness. For ignorable (MAR, MCAR) missingness, all scores within one covariate class had equal probability of being missing. Given these probabilities, a random sample of item scores was removed from the original data matrix.

Non-ignorable missingness (NMAR) was simulated as follows. Let $\overline{\psi}_{jq}$ be the mean location parameter of item *j* (equation (7)). Within one covariate class, for items with $\overline{\psi}_{jq} \geq 0$ scores of $X_{ij} \geq 3$ had a higher probability of being missing than smaller scores: For $\overline{\psi}_{jq} = 0$ this probability was twice as high, for $\overline{\psi}_{jq} = 1$ this probability was four times as high, and for $\overline{\psi}_{jq} = 2$ this probability was six times as high. This type of missingness may occur when people with higher latent variable values are reluctant to answer questions that may reveal their latent variable value.

Together with the influence of the covariate this manipulation of the response probabilities resulted in four different missingness mechanisms. Missingness was MCAR if it depended neither on *Y* nor on X_{ij} and $\overline{\psi}_{jq}$; MAR if it depended only on *Y*; NMAR of type 1 [denoted NMAR(1)] when it depended on X_{ij} and $\overline{\psi}_{jq}$; and NMAR(2) when it depended on *Y*, X_{ij} and $\overline{\psi}_{jq}$; see Table 3. Table 4 shows the probability ratios for the four missingness mechanisms, all values of *Y*, all values of the mean location parameter $\overline{\psi}_{jq}$, and of all values of X_{ij}.

Table 3. The relation of the four different missingness mechanisms used in the simulation study: MCAR, MAR, NMAR(1) and NMAR(2)

	Is missingness related to value of *Y*?	
	No	Yes
Is missingness related to values of X_{ij} and $\overline{\psi}_{jq}$?		
No	MCAR	MAR
Yes	NMAR(1)	NMAR(2)

Ignoring the covariate. Ignoring a relevant covariate produces NMAR (Schafer, 1997, p. 23). Its effect was compared to properly taking the covariate into account. For method TW-E, the influence of the covariate was evaluated by using TW-E in both classes separately. For methods TW-SS_{tw}, TW-FL_{tw}, TW-SS_{bs} and TW-FL_{bs} scores were first imputed for both classes separately using method TW-E. Next, a PCA/VR solution was obtained for the whole completed data and then scores were imputed in the incomplete data for both classes separately, using for both classes the same PCA/VR results. For methods TW-SS_{od} and TW-FL_{od}, a PCA/VR solution was obtained for the original data after which the imputation methods were used in both classes separately, using for both classes the same PCA/VR results. For method MNI, the covariate was included in the multivariate normal model that was estimated from the data. Ignoring the covariate meant that each of the five TW methods was used for imputation in the whole data set, and for method MNI the covariate was not included in the multivariate normal model estimated from the data.

Table 4. Probability ratios for all missingness mechanisms, and all values of covariate *Y*, mean location parameter $\overline{\psi}_{jq}$, and item score X_{ij}

		Y									
		1					2				
		X_{ij}					X_{ij}				
Missingness mechanism	$\overline{\psi}_{jq}$	0	1	2	3	4	0	1	2	3	4
MCAR	−2	1	1	1	1	1	1	1	1	1	1
	−1	1	1	1	1	1	1	1	1	1	1
	0	1	1	1	1	1	1	1	1	1	1
	1	1	1	1	1	1	1	1	1	1	1
	2	1	1	1	1	1	1	1	1	1	1
MAR	−2	1	1	1	1	1	2	2	2	2	2
	−1	1	1	1	1	1	2	2	2	2	2
	0	1	1	1	1	1	2	2	2	2	2
	1	1	1	1	1	1	2	2	2	2	2
	2	1	1	1	1	1	2	2	2	2	2
NMAR(1)	−2	1	1	1	1	1	1	1	1	1	1
	−1	1	1	1	1	1	1	1	1	1	1
	0	1	1	1	2	2	1	1	1	2	2
	1	1	1	1	4	4	1	1	1	4	4
	2	1	1	1	6	6	1	1	1	6	6
NMAR(2)	−2	1	1	1	1	1	2	2	2	2	2
	−1	1	1	1	1	1	2	2	2	2	2
	0	1	1	1	2	2	2	2	2	4	4
	1	1	1	1	4	4	2	2	2	8	8
	2	1	1	1	6	6	2	2	2	12	12

2.3.3. Dependent variables

Cronbach's alpha is reported in almost every study that uses tests or questionnaires; Loevinger's *H* is an easy-to-use coefficient that evaluates the scalability of a set of items (see Sijtsma and Molenaar, 2002, for an overview of approximately 30 applications); and Mokken's (1971) item selection cluster-algorithm is used for investigating the dimensionality of test and questionnaire data (see, for example, Van Abswoude, Van der Ark, & Sijtsma, 2004). These three dependent variables provide a good impression of the degree of success of the proposed imputation methods.

Results of Cronbach's alpha and coefficient H. Definitions and computations concerning Cronbach's alpha and coefficient *H* run parallel; thus, to avoid redundancy we focus exclusively on alpha. First, the *discrepancy* between Cronbach's alpha based on completed data and Cronbach's alpha based on corresponding original data was computed. Second, the *bias* of Cronbach's alpha based on completed data relative to the population value was computed. Note that there were four population values of Cronbach's alpha, one for each item subset that was driven by a particular θ_q ($q = 1, \ldots, 4$), and denoted α_q. Bias was computed for each item subset.

Computations were done as follows. Cronbach's alpha was computed for each item subset in each original data set (indexed $v = 1, \ldots, 100$), and denoted $\hat{\alpha}_{\text{or},vq}$; and for

each of the five completed data sets corresponding to original data set v. The mean of these five values was denoted $\hat{\alpha}_{\text{imp},vq}$. Bias in the original data was $\hat{\alpha}_{\text{or},vq} - \alpha_q$ and bias in the completed data was $\hat{\alpha}_{\text{imp},vq} - \alpha_q$. Discrepancy in alpha was defined as $\hat{\alpha}_{\text{imp},vq} - \hat{\alpha}_{\text{or},vq}$. Both bias and discrepancy served as dependent variables in ANOVAs. The tables contain the mean (*M*) and the standard deviation (*SD*) of the bias/discrepancy calculated over 100 replications.

For method LD, alpha was computed for the available complete cases and denoted $\hat{\alpha}_{\text{cc},vq}$. Bias and discrepancy produced by method LD are defined as $\hat{\alpha}_{\text{cc},vq} - \alpha_q$ and $\hat{\alpha}_{\text{cc},vq} - \hat{\alpha}_{\text{or},vq}$, respectively. Because each item subset produced a bias/discrepancy estimate, 'item subset' was included as a within-subjects factor in ANOVA.

Results of cluster solution from Mokken scale analysis. In exploratory Mokken scale analysis, one or more scales are selected from the data using a sequential cluster algorithm, described in detail by Mokken (1971) and Sijtsma and Molenaar (2002). The program MSP (Molenaar & Sijtsma, 2000) was used for this analysis. First, by assigning the items to the clusters in which they were selected most frequently the modal cluster solution was determined for the five completed data sets; see Van der Ark and Sijtsma (2005) for details. Second, the minimum number of items to be moved from the modal cluster solution to reobtain the four theoretical scales from the simulation model was determined, and called the *population classification error*. Also, the minimum number of items to be moved from the modal cluster solution to reobtain the original-data cluster solution was computed, and called the *original-data classification error*.

For method LD, the classification errors were the minimum number of items to be moved from the cluster solution based on the available complete cases in order to re-obtain the population cluster solution and the original-data cluster solution, respectively. The means (*M*) and standard deviations (*SD*) of the classification errors across 100 replicated data sets are reported.

2.3.4. Main design

The seven independent variables were: (1) Correlation between latent variables ($\rho = 0$, .24, .5); (2) sample size ($N = 300$, 1,000); (3) percentage of missingness (1%, 5%, 15%); (4) effect of covariate on missingness (No, Yes); (5) joint effect of item score and location parameters on missingness (No, Yes); (6) ignoring covariate (No, Yes); and (7) missing-data method (LD, TW-E, TW-SS_{tw}, TW-FL_{tw}, TW-SS_{od}, TW-FL_{od} and MNI). An item had five answer categories and the number of items was 40 (see Table 5 for the design characteristics).

2.3.5. Three specialized designs

Specialized design: Unequal correlations between latent variables. In practice, the correlations between latent variables are likely to be unequal. Thus, in a specialized design the correlation matrix of the latent variables was

$$\sum = \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ .24 & .50 & 1 & \\ .50 & .24 & 0 & 1 \end{bmatrix},$$

Table 5. Independent variables and fixed characteristics of the main design

Independent variables	Levels
Correlation between latent variables	0, .24, .50
Sample size	300, 1000
Percentage of missingness	1, 5, 15
Effect of covariate on missingness	No, Yes
Joint effect of item scores and location parameters on missingness	No, Yes
Ignoring covariate	No, Yes
Imputation methods	TW-E, TW-FL_{tw}, TW-SS_{tw}, TW-FL_{od}, TW-SS_{od}, MNI
Fixed design characteristics	**Value**
Number of latent variables	4; multivariate normally distributed
Number of items	40
Number of answer categories	5
Number of imputations	5
Item parameters	Fixed per item, see Tables 1 and 2

using correlations from the main design. Methods TW-E, TW-SS_{tw}, TW-FL_{tw}, TW-SS_{od}, TW-FL_{od} and MNI were studied, the sample size was fixed at $N = 1,000$, the percentage of missingness was 5%, the missingness mechanism was MAR, and the effect of the covariate was taken into account.

Specialized design: Confidence intervals. Because MI corrects confidence intervals of parameter estimates using Rubin's (1987) rules, this specialized design studied focused on this topic. Kristof's (1963) derivation of the sampling distribution of Cronbach's alpha assumes multivariate normality and compound symmetry. Thus, data were sampled from a multivariate standard normal distribution for the item scores ($J = 40$). Even though these assumptions do not hold for highly discrete questionnaire data, Kristof's results were used as a benchmark for performance of the multiple-imputation methods.

There were four scales, and each scale consisted of 10 items. Items within the same scale correlated .5 and items from different scales correlated 0. There was no covariate. Methods TW-E, TW-SS_{tw}, TW-FL_{tw}, TW-SS_{od}, TW-FL_{od} and MNI were used, sample size was 1,000, missingness mechanism was MCAR and percentage of missingness was 5. One thousand replications were drawn to have more accurate estimates of the confidence intervals.

Because Rubin's rules for MI are defined for normally distributed variables, the non-normal sampling distribution of Cronbach's alpha was transformed into an approximately normal Fisher z-score by means of

$$z = \frac{1}{2} \ln \left(\frac{1}{1 - \hat{\alpha}} \right)$$

(e.g. McGraw & Wong, 1996). The number of replicated data sets out of 1,000 in which α_q was covered by the confidence interval was counted, and the mean (*M*) and standard deviation (*SD*) of the bias were computed.

The sampling distribution of coefficient H has been derived only for binary items (Mokken, 1971, pp. 157–169). Thus, confidence intervals for coefficient H were not

considered here. Mokken scale analysis was not considered because its outcome is not a parameter estimate.

Specialized design: Bootstrap methods. Bootstrap methods TW-SS_{bs} and TW-FL_{bs} were compared with methods TW-SS_{tw}, TW-FL_{tw}, TW-SS_{od} and TW-FL_{od}. It was expected that the bootstrap would be more relevant for large percentages of missingness; thus, 15% missingness was studied here. Missingness mechanism was MAR and covariate was taken into account. Correlations between latent variables were 0, .24 and .50, and sample sizes were 300 and 1,000; for this design choice the largest differences between the bootstrap and the other methods were expected.

2.4. Statistical analyses

ANOVAs were used to analyze bias and discrepancy in coefficients alpha and H. Sample size was treated as a between-subjects factor. All other factors were dependent measures and treated as within-subjects factors. Because classification error is discrete and skewed, a logistic regression with binomial counts was used. Let y_{vt} be the classification error of data set v in within-subjects design cell t, and e_{vt} the maximum number of items that are incorrectly clustered. For a test of 40 items, we have $e_{vt} = 39$. Let $\boldsymbol{\beta}$ be a column vector with regression coefficients, and for simulated data set v let $\mathbf{z}_v$ be a row vector with responses to the independent (dummy) variables. The probability that one item is incorrectly clustered is modelled as

$$\pi_{t,\mathbf{z}_v} = \frac{\exp(\mathbf{z}_v\boldsymbol{\beta})}{1 + \exp(\mathbf{z}_v\boldsymbol{\beta})}.$$

The logistic regression model with binomial counts is

$$P(y_{vt}|\mathbf{z}_v, e_{vt}) = \left[\frac{e_{vt}!}{y_{vt}(e_{vt} - y_{vt})!}\right](\pi_{t,\mathbf{z}_v})^{y_{vt}}(1 - \pi_{t,\mathbf{z}_v})^{e_{vt}-y_{vt}}.$$

(Vermunt & Magidson, 2005a, p. 11). Sample size was treated as an independent measure and the other factors as dependent measures. The logistic regression analyses with binomial counts were done using Latent Gold 4.0 (Vermunt & Magidson, 2005b).

3. Results

Both for alpha and H, the standard deviations of bias were approximately 10 times larger than the standard deviations of discrepancy. Because the original data produced unbiased estimates of alpha and H in all situations (one-sample t tests) the mean bias and the mean discrepancy were almost identical. Thus, it is sufficient to discuss only bias and ignore discrepancy.

The results for the population classification error in Mokken scale analysis deviated substantially from those for the original-data classification error. However, this was not entirely an effect related to missing-data problems but also of MSP having trouble finding the population cluster solution when correlations between latent variables were relatively high (see Van Abswoude *et al.*, 2004, for similar conclusions). Because the modal cluster solutions for the completed data and those for the corresponding original data were often similar, it made sense to study only the original-data classification error, henceforth called the classification error for short.

The results of method LD were much worse than the results of MI methods. LD produced bias with larger standard deviations, relatively large bias for 5% missingness and under departures from MCAR bias increased dramatically while bias due to MI methods was much smaller. For 15% missingness, almost no complete cases were available. Therefore, results of method LD are not further discussed.

3.1. Results of main design

3.1.1. Bias in Cronbach's alpha

A full-factorial ANOVA was conducted on the data from the completely crossed design of order 4 (item subsets) × 2 (sample size) × 3 (correlation) × 2 (percentage of missingness) × 2 (effect of covariate) × 2 (effect of nonignorable missingness) × 2 (ignoring covariate) × 6 (imputation method), with bias in Cronbach's alpha as dependent variable. One hundred and nine effects out of 240 were statistically significant, but Table 6 only shows the small, medium and large effects and their effect sizes (based on Cohen, 1988). Table 7 shows that bias was usually small; it ranged from −0.049 (method TW-E, 15% missingness, $\rho = 0$) to 0 (method TW-SS_{tw}, 1% missingness, $\rho = 0$) (Table 7). Standard deviations ranged from 0.011 to 0.024.

Table 6. Effect size of ANOVA with bias in Cronbach's alpha as dependent variable

Effect	*F*	df_1	df_2	η^2
Method	53844.39	5	990	.16***
Percentage of missingness	21819.53	2	396	.08**
Method × percentage of missingness	46632.58	10	1980	.12**
Method × correlation	33067.88	10	1980	.02*

*Small effect; **medium effect; ***large effect.
All *p*-values smaller than .001.

Imputation method × percentage missingness. For all imputation methods, bias in Cronbach's alpha increased as percentage of missingness increased (Table 7), but for 1% missingness bias was small. Increase in bias was larger for methods that produced larger bias for 1% missingness.

Imputation method × correlation. Bias produced by method MNI and the TW-SS methods was the same for different correlations (Table 7). Bias produced by the TW-FL methods increased little and for method TW-E bias decreased as the correlation between latent variables increased (Table 7).

Percentage of missingness. As the percentage of missingness increased, the bias in Cronbach's alpha also increased (Table 7).

Imputation method. Biases in Cronbach's alpha due to methods TW-FL_{tw}, TW-FL_{od}, and method MNI were similar. Methods TW-SS_{tw} and TW-SS_{od} produced smaller bias and method TW-E the largest bias (Table 7).

3.1.2. Bias in coefficient H

Means and standard deviations of bias in *H* were somewhat larger than those for Cronbach's alpha, but conclusions were similar (Table 8). Bias ranged from −0.082 (method TW-E, 15% missingness, $\rho = 0$) to 0 (method TW-FL_{tw}, 1% missingness, $\rho = 0$). Standard deviations of bias ranged from 0.019 to 0.039.

Table 7. Mean (*M*) and standard deviation (*SD*) of bias in Cronbach's alpha for all combinations of percentage of missingness, imputation method, and correlation between latent variables

		Correlation between latent variables							
		0		.24		.50		Mean	
Percentage of missingness	Imputation method	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1%	TW-SS_{tw}	0	11	0	11	0	11	0	11
	TW-FL_{tw}	0	11	−1	11	−1	11	−1	11
	TW-SS_{od}	0	11	0	11	0	11	0	11
	TW-FL_{od}	0	11	−1	11	−1	11	−1	11
	TW-E	−3	11	−3	11	−2	11	−3	11
	MNI	0	11	−1	11	−1	11	−1	11
	Mean	−1	11	−1	11	−1	11	−1	11
5%	TW-SS_{tw}	1	11	0	11	0	11	0	11
	TW-FL_{tw}	−2	11	−4	11	−4	11	−3	11
	TW-SS_{od}	0	11	0	11	0	11	0	11
	TW-FL_{od}	−2	11	−3	11	−4	11	−3	11
	TW-E	−15	12	−12	11	−9	12	−12	12
	MNI	−2	11	−2	11	−2	11	−2	11
	Mean	−3	13	−3	12	−3	12	−3	12
15%	TW-SS_{tw}	2	11	1	11	1	11	1	11
	TW-FL_{tw}	−10	13	−13	14	−14	14	−12	14
	TW-SS_{od}	1	11	1	11	1	11	1	11
	TW-FL_{od}	−6	12	−10	13	−12	13	−9	13
	TW-E	−49	17	−37	15	−16	14	−37	18
	MNI	−7	12	−7	12	−7	12	−7	12
	Mean	−11	22	−11	18	−11	18	−9	16
Mean	TW-SS_{tw}	1	11	1	11	1	11	1	11
	TW-FL_{tw}	−4	17	−4	13	−4	13	−4	13
	TW-SS_{od}	1	11	1	11	1	11	1	11
	TW-FL_{od}	−3	12	−3	12	−3	12	−3	12
	TW-E	−22	24	−22	24	−22	24	−22	24
	MNI	−3	12	−3	12	−3	12	−3	12

Entries in table must be multiplied by 10^{-3}.

3.1.3. Classification error

A full-factorial logistic regression with binomial counts was conducted on a completely crossed design of order 2 (sample size) × 3 (correlation) × 2 (percentage of missingness) × 2 (effect of covariate) × 2 (effect of non-ignorable missingness) × 2 (ignoring covariate) × 6 (imputation method), with classification error as dependent variable. Thirty-seven effects out of 127 were significant. Only the largest effects are reported. Table 9 shows classification error results for all combinations of imputation method and correlation between latent variables.

Imputation method × correlation between latent variables. The interaction effect of imputation method and correlation between latent variables was significant (Wald test; $\chi^2(10) = 874.38$, $p < .001$). For $\rho = 0$ and .24, classification error was smaller than for $\rho = .50$ (Table 9). For methods TW-FL_{od} and TW-FL_{tw}, the effect of correlation

Table 8. Mean (*M*) and standard deviation (*SD*) of bias in coefficient *H* for all combinations of percentage of missingness, imputation method and correlation between latent variables

		Correlation between latent variables							
		0		.24		.50		Mean	
Percentage of missingness	Imputation method	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1%	TW-SS_{tw}	2	20	1	19	1	20	1	20
	TW-FL_{tw}	1	20	0	19	0	20	0	20
	TW-SS_{od}	2	20	2	19	1	20	1	20
	TW-FL_{od}	1	20	0	19	0	20	0	20
	TW-E	−4	20	−3	19	−2	20	−2	20
	MNI	1	20	1	19	1	20	1	20
	Mean*	0	20	0	19	0	20	0	20
5%	TW-SS_{tw}	1	20	0	19	0	20	0	19
	TW-FL_{tw}	−4	20	−7	19	−7	20	−6	20
	TW-SS_{od}	1	20	0	19	0	20	0	19
	TW-FL_{od}	−3	20	−6	19	−7	20	−6	20
	TW-E	−27	20	−22	19	−16	20	−22	20
	MNI	−2	20	−3	20	−3	20	−3	20
	Mean	−6	22	−6	21	−6	20	−6	21
15%	TW-SS_{tw}	−2	20	−3	20	−4	21	−3	21
	TW-FL_{tw}	−21	24	−27	24	−27	24	−25	24
	TW-SS_{od}	−3	20	−3	20	−3	21	−3	20
	TW-FL_{od}	−15	22	−22	22	−24	23	−20	23
	TW-E	−82	23	−65	22	−47	21	−65	26
	MNI	−13	21	−13	21	−13	21	−13	21
	Mean	−22	35	−22	30	−20	27	−21	31
Mean	TW-SS_{tw}	−2	20	0	20	−1	20	0	20
	TW-FL_{tw}	−10	23	−11	24	−12	24	−10	24
	TW-SS_{od}	−2	20	0	20	−1	20	0	20
	TW-FL_{od}	−8	22	−9	22	−10	23	−8	22
	TW-E	−40	39	−30	33	−22	28	−30	34
	MNI	−7	21	−9	21	−5	21	−5	21

Entries in table must be multiplied by 10^{-3}.

resembled that for method MNI. Methods TW-SS_{tw} and TW-SS_{od} produced the largest classification error for $\rho = .50$, but method TW-E produced nearly the same classification error for different correlations.

Imputation method × percentage of missingness. The interaction of imputation method and percentage of missingness was significant ($\chi^2(10) = 732.16$, $p < .001$). Classification error increased as percentage of missingness increased and methods that produced a relatively large classification error for 1% missingness also produced a larger classification error as percentage of missingness increased further (Table 9).

Imputation method. A significant effect of imputation method was found, ($\chi^2(5) = 1562.64$, $p < .001$). Methods MNI, TW-FL_{tw} and TW-FL_{od} produced the smallest classification error Method TW-E produced the largest classification error.

Table 9. Mean (*M*) and standard deviation (*SD*) of classification error for all combinations of imputation method, percentage of missingness and correlation between latent variables

		Correlation between latent variables							
		0		.24		.50		Mean	
Percentage of missingness	Imputation method	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1%	TW-SS_{tw}	0.45	0.77	0.53	0.85	1.59	2.64	0.86	1.74
	TW-FL_{tw}	0.48	0.79	0.56	0.85	1.47	2.50	0.83	1.65
	TW-SS_{od}	0.44	0.75	0.54	0.85	1.59	2.78	0.86	1.81
	TW-FL_{od}	0.46	0.77	0.55	0.85	1.43	2.34	0.81	1.57
	TW-E	0.71	0.90	0.73	0.98	1.48	2.46	0.97	1.66
	MNI	0.52	0.81	0.64	1.00	1.37	2.36	0.84	1.59
	Mean	0.51	0.80	0.59	0.90	1.49	2.52	0.86	1.67
5%	TW-SS_{tw}	0.99	1.04	1.14	1.31	4.09	4.56	2.07	3.15
	TW-FL_{tw}	0.98	1.05	1.25	1.35	2.71	3.09	1.65	2.17
	TW-SS_{od}	0.99	1.04	1.13	1.25	4.20	4.76	2.11	3.26
	TW-FL_{od}	0.99	1.05	1.19	1.25	2.78	3.21	1.65	2.23
	TW-E	2.60	1.35	2.27	1.56	2.71	3.06	2.52	2.14
	MNI	1.19	1.17	1.33	1.34	2.55	3.11	1.69	2.16
	Mean	1.29	1.27	1.38	1.41	3.17	3.77	1.95	2.58
15%	TW-SS_{tw}	1.89	1.23	2.01	1.53	15.89	8.94	6.60	8.44
	TW-FL_{tw}	2.23	1.49	2.73	1.79	4.88	4.34	3.28	3.06
	TW-SS_{od}	1.86	1.24	2.01	1.54	15.90	8.86	6.59	8.42
	TW-FL_{od}	1.91	1.38	2.29	1.70	4.94	4.47	3.05	3.17
	TW-E	8.49	1.93	6.68	2.10	3.71	3.38	6.29	3.23
	MNI	2.37	1.46	2.55	1.75	3.82	3.53	2.91	2.51
	Mean	3.12	2.82	3.04	2.40	8.19	8.17	4.79	5.71
Mean	TW-SS_{tw}	1.11	1.19	1.23	1.40	7.19	8.65	3.17	5.84
	TW-FL_{tw}	1.23	1.36	1.51	1.65	3.02	3.68	1.92	2.58
	TW-SS_{od}	1.10	1.18	1.22	1.39	7.23	8.66	3.18	5.88
	TW-FL_{od}	1.12	1.25	1.34	1.50	3.05	3.74	1.83	2.58
	TW-E	3.93	3.62	3.23	3.00	2.63	3.13	3.26	3.30
	MNI	1.33	1.40	1.47	1.62	2.58	3.16	1.79	2.27
	Mean	1.64	2.15	1.67	1.98	4.28	6.10	2.53	4.09

Correlation between latent variables. Correlation between latent variables had a significant main effect ($\chi^2(2) = 204.82, p < .001$). For $\rho = 0$ and .24, classification error was smaller than for $\rho = .50$ (Table 9).

Percentage of missingness. Percentage of missingness had a significant effect (Wald test; $\chi^2(2) = 2418.22$, $p < .001$). As percentage of missingness increased, the magnitude of the classification error also increased (Table 9).

3.2. Results of specialized designs

Unequal correlations between latent variables. An ANOVA was conducted on a completely crossed design of order 4 (item subset) × 4 (correlation) × 6 (imputation method). For Cronbach's alpha, two effects were found: a small interaction effect of correlation between latent variables and imputation method ($F(15, 1485) = 8470.36$, $p < .001, \eta^2 = .01$) and a large main effect of imputation method ($F(5, 495) = 18{,}955.68$, $p < .001$, $\eta^2 = .22$). For coefficient *H*, again two effects were found: a small

interaction effect of imputation method and correlation between latent variables ($F(15, 1485) = 7994.57$, $p < .001$, $\eta^2 = .01$) and a large main effect of imputation method ($F(5, 495) = 19180.15, p < .001, \eta^2 = .22$).

For classification error, a logistic regression with binomial counts on the 4 (correlation) × 6 (imputation method) design showed significant effects for correlation between latent variables ($\chi^2(3) = 289.21$, $p < .001$), imputation method ($\chi^2(5) = 135.29$, $p < .001$) and interaction of imputation method and correlation between latent variables ($\chi^2(15) = 210.90$, $p < .001$).

Table 10 shows that for unequal correlations between latent variables, bias in Cronbach's alpha and coefficient *H* was similar to bias for equal correlations. Unequal correlations produced the largest classification error for all imputation methods.

Confidence intervals. Table 11 shows that methods TW-FL_{tw} and MNI closely recover the theoretical 95% confidence intervals for Cronbach's alpha (upper panel), directly followed by method TW-SS_{tw}. Method TW-E performs worst: Only 60% of the simulated confidence intervals cover α_q. Furthermore, method MNI produces the smallest bias, followed by method TW-FL_{tw} (bottom panel). This result is due to MNI assuming multivariate normal data, which is the data model used here. Method TW-E produces the greatest bias.

Bootstrap methods. Imputation method had a large main effect on both bias in Cronbach's alpha and coefficient *H* (Cronbach's alpha: $F(5, 990) = 11,754.49$, $p < .001$, $\eta^2 = .19$; coefficient *H*: $F(5, 990) = 10,516.80$, $p < .001$, $\eta^2 = .17$). Biases produced by methods TW-SS_{tw}, TW-SS_{od} and TW-SS_{bs} were equally large (Table 12). Method TW-FL_{tw} produced a slightly smaller bias in Cronbach's alpha and coefficient *H* than method TW-FL_{bs}. Of the TW-FL methods, method TW-FL_{od} produced the smallest bias.

All effects on classification error in Mokken scale analysis were significant; means and standard deviations are reported in Table 13. Differences in classification error among methods TW-FL_{tw}/TW-SS_{tw}, methods TW-FL_{od}/TW-SS_{od} and methods TW-FL_{bs}/TW-SS_{bs} were small (Table 13).

4. Discussion

Bias produced by multiple-imputation versions of variations on method Two-way is mainly influenced by percentage of missingness and correlation between latent variables. Thus, a good MI method should be robust against variations of these factors. Sample size and missingness mechanism were not as influential.

Method TW-SS_{tw} is the preferred method because it produced almost no bias in Cronbach's alpha and coefficient *H* for various percentages of missingness and correlations between latent variables. Method MNI also produced small bias but, in general, was outperformed by method TW-SS_{tw}.

For Mokken scale analysis, method TW-FL_{tw} is a better alternative. It produced the smallest classification error for $\rho = .50$. However, as long as the different item clusters in the data are not highly correlated, method TW-SS_{tw} may also be used.

One noticeable result was that methods that used factor loadings from the original data (i.e. methods TW-SS_{od} and TW-FL_{od}) did not produce smaller bias than methods that used factor loadings from a completed data set using method TW-E (methods TW-SS_{tw} and TW-FL_{tw}). Moreover, methods that estimated the factor loadings from completed bootstrap data sets (TW-FL_{bs} and TW-SS_{bs}) did not produce smaller bias and sometimes produced even larger bias than methods that estimated factor loadings from the completed data set (TW-FL_{tw} and TW-SS_{tw}). Thus, from a practical point of view,

Table 10. Mean (*M*) and standard deviation (*SD*) of bias in Cronbach's alpha, coefficient *H*, and the classification error, for specialized design with unequal correlation between latent variables

		Correlation between latent variables									
		0		.24		.50		Unequal		Mean	
Dependent variable	Method	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Bias in alpha	TW-SS_{tw}	1	8	1	7	1	8	1	8	1	8
	TW-FL_{tw}	−1	8	−3	7	−3	8	−2	8	−2	8
	TW-SS_{od}	1	8	1	7	1	8	1	8	1	8
	TW-FL_{od}	−1	8	−2	7	−3	8	−2	8	−2	8
	TW-E	−15	8	−11	8	−8	8	−11	8	−11	8
	MNI	−2	8	−2	8	−2	8	−1	8	−2	8
Bias in *H*	TW-SS_{tw}	0	14	0	13	0	14	1	14	0	14
	TW-FL_{tw}	−3	14	−6	14	−6	14	−5	14	−5	14
	TW-SS_{od}	0	14	0	13	0	14	1	14	0	14
	TW-FL_{od}	−2	14	−5	13	−6	14	−4	14	−4	14
	TW-E	−27	13	−22	13	−15	13	−20	13	−21	14
	MNI	−3	14	−3	14	−3	14	−2	14	−3	14
Classification error	TW-SS_{tw}	0.85	0.82	0.86	0.93	4.25	5.81	6.62	5.88	3.14	4.83
	TW-FL_{tw}	0.71	0.86	0.74	0.84	1.71	2.74	3.52	3.63	1.67	2.61
	TW-SS_{od}	0.88	0.90	0.89	0.86	4.49	6.79	7.07	5.62	3.33	5.15
	TW-FL_{od}	0.82	0.88	0.78	0.91	1.95	3.66	3.63	3.85	1.80	2.96
	TW-E	2.46	1.25	1.90	1.29	1.99	3.66	4.87	3.79	2.80	3.02
	MNI	0.84	0.91	0.88	0.96	1.84	3.65	3.66	3.83	1.81	2.95

Entries in upper and middle panel must be multiplied by 10^{-3}.

Table 11. Simulated 95% confidence intervals and mean (*M*) and standard deviation (*SD*) of bias in Cronbach's alpha for specialized design with multivariate normally distributed variables

		Scale							
	Method	1		2		3		4	
Confidence intervals of Cronbach's alpha	Original data	94.5		94.0		94.3		95.4	
	TW-SS_{tw}	92.6		90.9		94.3		93.6	
	TW-FL_{tw}	95.0		93.5		95.4		94.7	
	TW-SS_{od}	93.1		91.5		94.5		94.0	
	TW-FL_{od}	94.8		93.6		94.8		94.5	
	TW-E	61.2		61.8		58.3		58.0	
	MNI	94.7		94.2		94.0		95.6	
	Method	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Bias in Cronbach's alpha	Original data	0	9	0	10	−1	9	0	9
	TW-SS_{tw}	4	9	4	9	3	9	4	9
	TW-FL_{tw}	2	9	2	9	1	9	1	9
	TW-SS_{od}	4	9	4	9	3	9	4	9
	TW-FL_{od}	2	9	2	9	1	9	2	9
	TW-E	−16	10	−16	10	−17	10	−17	9
	MNI	0	9	0	10	−1	9	−1	9

Entries in bottom panel must be multiplied by 10^{-3}.

methods TW-FL_{tw} and TW-SS_{tw} may be preferred over methods TW-FL_{bs} and TW-SS_{bs}, even though the latter may be argued to be theoretically superior.

Another noticeable result was that when scores were MAR, including the covariate in the analysis had little effect on bias. This result may have been due to the factorial structure being the same in both covariate classes while the different latent variable means may have been too close to make a difference. NMAR mechanisms did not have a discernable effect on the bias in Cronbach's alpha and coefficient *H* either, but an effect was found for discrepancy (results not discussed).

Table 12. Mean (*M*) and standard deviation (*SD*) of bias in Cronbach's alpha and coefficient *H*, for specialized design with methods TW-FL_{bs} and TW-SS_{bs} in addition to four other TW methods

	Bias			
	α		*H*	
Imputation method	*M*	*SD*	*M*	*SD*
TW-SS_{tw}	3	11	1	20
TW-FL_{tw}	−8	12	−17	20
TW-SS_{od}	3	11	1	19
TW-FL_{od}	−6	12	−13	20
TW-SS_{bs}	3	11	1	20
TW-FL_{bs}	−9	12	−19	19

Entries must be multiplied by 10^{-3}.

Table 13. Mean (*M*) and standard deviation (*SD*) of classification error for all combinations of sample size, imputation method and correlation between latent variables, for specialized design with methods TW-SS_{bs} and TW-FL_{bs}, in addition to four other TW methods

		Correlation between latent variables							
		0		.24		.50		Mean	
Sample size	Imputation method	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
300	TW-SS_{tw}	2.36	1.34	2.55	1.73	10.79	6.04	5.23	5.40
	TW-FL_{tw}	2.23	1.38	2.50	1.78	5.25	3.49	3.33	2.75
	TW-SS_{od}	2.38	1.41	2.67	1.86	11.08	5.69	5.38	5.37
	TW-FL_{od}	2.14	1.42	2.34	1.77	5.23	2.98	3.24	2.58
	TW-SS_{bs}	2.27	1.41	2.49	1.80	10.90	6.13	5.22	5.51
	TW-FL_{bs}	2.20	1.29	2.66	1.85	4.87	3.38	3.24	2.61
	Mean	2.26	1.37	2.54	1.80	8.02	5.60	4.27	4.38
1000	TW-SS_{tw}	1.92	1.02	1.79	1.09	21.19	9.07	8.30	10.55
	TW-FL_{tw}	1.60	1.13	1.50	1.13	3.29	4.52	2.13	2.88
	TW-SS_{od}	1.94	1.10	1.72	1.09	20.59	8.96	8.08	10.29
	TW-FL_{od}	1.66	1.03	1.38	1.07	3.29	3.96	2.11	2.58
	TW-SS_{bs}	1.94	1.18	1.89	1.14	21.23	9.36	8.35	10.64
	TW-FL_{bs}	1.54	0.97	1.56	1.07	3.07	4.19	2.06	2.65
	Mean	1.77	1.08	1.64	1.11	12.11	11.38	5.17	8.25
Mean	TW-SS_{tw}	2.14	1.21	2.17	1.49	15.99	9.29	6.77	8.51
	TW-FL_{tw}	1.92	1.29	2.00	1.57	4.27	4.14	2.73	2.88
	TW-SS_{od}	2.16	1.28	2.20	1.60	15.84	8.88	6.73	8.31
	TW-FL_{od}	1.90	1.26	1.86	1.54	4.26	3.63	2.67	2.64
	TW-SS_{bs}	2.11	1.31	2.19	1.53	16.07	9.44	6.79	8.61
	TW-FL_{bs}	1.87	1.18	2.11	1.60	3.97	3.90	2.65	2.69
	Mean	2.01	1.26	2.09	1.56	10.06	9.20	4.72	6.62

To summarize, method TW-SS_{tw} in particular, and method TW-FL_{tw} are promising and simpler alternatives to method MNI for multidimensional rating-scale test and questionnaire data. These methods are easily accessible in SPSS (subroutines due to Van Ginkel & Van der Ark, 2005). Method MNI is applicable in many missing-data problems and may be the preferred method for many researchers who are used to it already, also when the data are multi-item, multidimensional and highly discrete. Nevertheless, in psychometric work the simple method TW-SS_{tw} is a good alternative.

Acknowledgements

The authors would like to thank Jeroen Vermunt for assistance with the analyses in Latent Gold 4.0.

References

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research*, *35*, 321–364.

Brough, P., O'Driscoll, M., & Kalliath, T. (2005). Confirmatory factor analysis of the Cybernetic coping scale. *Journal of Occupational and Organizational Psychology*, *78*, 53–61.

Brown, T. A., White, K. S., & Barlow, D. H. (2005). A psychometric reanalysis of the Albany Panic and Phobia Questionnaire. *Behaviour Research and Therapy*, *43*, 337–355.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cronbach, J. L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, *39*, 1–38.

Ezzati-Rice, T. M., Johnson, W., Khare, M., Little, R. J. A., Rubin, D. B., & Schafer, J. L. (1995). A simulation study to evaluate the performance of model-based multiple imputations in NCHS health examination surveys. In *Proceedings of the Annual Research Conference* (pp. 257–266). Washington, DC: Bureau of the Census.

Fay, R. E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, *81*, 354–365.

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1–29). Thousand Oaks, CA: Sage.

Harman, H. H. (1976). *Modern factor analysis*. Chicago: University of Chicago Press.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*, 475–492.

Hills, P., Francis, J. F., & Robbins, M. (2005). The development of the Revised Religious Life Inventory (RLI-R) by exploratory and confirmatory factor analysis. *Personality and Individual Differences*, *38*, 1389–1399.

Huisman, M. (1998). *Item nonresponse: Occurrence, causes, and imputation of missing answers to test items*. Leiden, The Netherlands: DSWO Press.

Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika*, *59*, 149–176.

Kristof, W. (1963). The statistical theory of stepped-up reliability when a test has been divided into several equivalent parts. *Psychometrika*, *28*, 221–238.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Little, R. J. A., & Su, H. L. (1989). Item nonresponse in panel surveys. In D. Kasprzyk, G. Duncan, & M. P. Singh (Eds.), *Panel surveys* (pp. 400–425). New York: Wiley.

Loevinger, J. (1948). The technique of homogeneous test compared with some aspects of scale analysis and factor analysis. *Psychological Bulletin*, *45*, 507–530.

Mata, I., Mataix-Cols, D., & Peralta, V. (2005). Schizotypal Personality Questionnaire – Brief: Factor structure and influence of sex and age in a nonclinical population. *Personality and Individual Differences*, *38*, 1183–1192.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton/Berlin: De Gruyter.

Molenaar, I. W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows*. Groningen, The Netherlands: IecProGAMMA.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: A latent variable approach to item-nonresponse in attitude scales. *Journal of the Royal Statistical Society, Series A*, *162*, 177–194.

Raaijmakers, Q. O. L. (1999). Effectiveness of different missing data treatments in surveys with Likert type data: Introducing the relative mean substitution approach. *Educational and Psychological Measurement*, *59*, 725–748.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman and Hall.
Schafer, J. L. (1998). NORM: Version 2.02 for Windows 95/98/NT. Retrieved 26 April, 2006, from http://www.stat.psu.edu/ ~ jls/misoftwa.html
Schafer, J. L., Ezzati-Rice, T. M., Johnson, W., Khare, M., Little, R. J. A., & Rubin, D. B. (1996). *The NHANES III multiple imputation project.* Proceedings of the Survey Research Methods Section of the American Statistical Association(pp. 28–37). Retrieved 26 April, 2006, from http://www.amstat.org/sections/srms/Proceedings/papers/1996_004.pdf
Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.
Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage.
Sijtsma, K., & Van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research*, *38*, 505–528.
Smits, N., Mellenbergh, G. J., & Vorst, H. C. M. (2002). Alternative missing data techniques to grade point average: Imputing unavailable grades. *Journal of Educational Measurement*, *39*, 187–206.
S-Plus 6 for Windows [Computer software]. (2001). Seattle, WA: Insightful Corporation.
Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–540.
Van Abswoude, A. A. H., Van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement*, *28*, 3–24.
Van der Ark, L. A., & Sijtsma, K. (2005). The effect of missing data imputation on Mokken scale analysis. In L. A. van der Ark, M. A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 147–166). Mahwah, NJ: Erlbaum.
Van Ginkel, J. R., & Van der Ark, L. A. (2005). TW.SPS, RF.SPS, CIMS.SPS, TW-SS.SPS, and TW-FL.SPS [computer code]. Retrieved 26 April, 2006, from http://www.uvt.nl/mto/software2.html
Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2007). *Multiple imputation of test and questionnaire data and influence on psychometric results. Multivariate Behavioral Research, 42*, 387–414.
Vermunt, J. K., & Magidson, J. (2005a). *Technical guide for Latent GOLD 4.0: Basic and advanced*. Belmont, MA.
Vermunt, J. K., & Magidson, J. (2005b). Latent GOLD 4.0 [Computer software]. Belmont, MA.
Yuan, Y. C. (2000). *Multiple imputation for missing data: Concepts and new development.* Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference (Paper No. 267). Cary, NC: SAS Institute. Retrieved April 26, 2006, from http://www.ats.ucla.edu/stat/sas/library/multipleimputation.pdf

Received 25 October 2005; revised version received 26 April 2006

Appendix

This appendix presents a justification for weighting the item scores with the factor loadings (equation (5)). The reversed item score is computed as

$$X_{ij}^{r} = x_{\max} + x_{\min} - X_{ij}.$$

First, Harman (1976, p. 169) showed that if the loading of factor k on item j is a_{jk}, then the loading of factor k on the reversed scored item is $-a_{jk}$. The scores of an item are

reversed, the loading on factor k is retained with opposite sign. Hence, reversing all items with negative loadings on factor k is a way to circumvent negative loadings in the denominator of equation (4). Equation (4) may then be written as

$$PM_{ik} = \frac{\sum_{j \in obs(i)} |a_{jk}| \times Z_{ij}}{\sum_{j \in obs(i)} |a_{jk}|} \quad \begin{cases} Z_{ij} = X_{ij} & \text{for} \quad a_{jk} > 0 \\ Z_{ij} = X_{ij}^{r} & \text{for} \quad a_{jk} < 0 \end{cases} \tag{A1}$$

Second, for computational convenience we transform item score X_{ij} into

$$X_{ij}^{*} = X_{ij} - x_{\text{mid}} \tag{A2}$$

Transformation of (A2) produces item scores such that $X_{ij}^{*} = -X_{ij}^{*r}$, and thus the following condition is satisfied:

$$a_{jk} \times X_{ij}^{*} = -a_{jk} \times X_{ij}^{*r}. \tag{A3}$$

Equation (A1) can now be written as

$$PM_{ik}^{*} = \frac{\sum_{j \in obs(i)} a_{jk} \times X_{ij}^{*}}{\sum_{j \in obs(i)} |a_{jk}|}. \tag{A4}$$

To obtain the correct value of PM_{ik}^{*}, x_{mid} must be added to equation (A4) which yields equation (5).